Lecture 3: Introducing Statistical Modeling

36-401, Fall 2018, Section B

1 Motivating

Let's start this off with a motivating example. We'll begin by loading some data which comes from the Bureau of Economic Analysis, on the economic output of cities in the U.S. (`http://www.bea.gov/regional/gdpmetro/`).

```
bea = read.csv("http://www.stat.cmu.edu/~larry/=stat401/bea-2006.csv")
```

For each city — more precisely, each "Metropolitan Statistical Area", which ignores legal divisions of cities and counties and instead is based on patterns of commuting — this records the name of the city, its population, its per-capita "gross metropolitan product" in 2006 (the total value of goods and services produced), and the share of the economy coming from four selected industries.

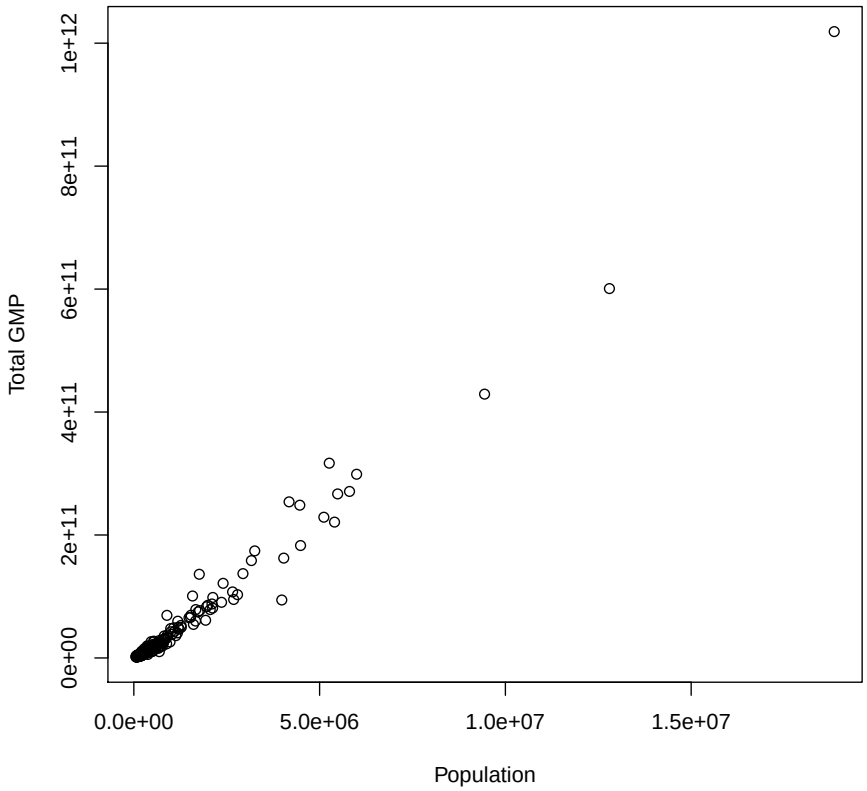
```
dim(bea)
## [1] 366   7
head(bea)
##                           MSA pcgmp    pop finance prof.tech     ict
## 1                 Abilene, TX 24490 158700 0.09750        NA 0.01621
## 2                   Akron, OH 32890 699300 0.12940   0.05440      NA
## 3                  Albany, GA 24270 163000 0.08217        NA 0.00708
## 4 Albany-Schenectady-Troy, NY 36840 850300 0.15780   0.09399 0.04511
## 5             Albuquerque, NM 37660 816000 0.15990   0.09978 0.20500
## 6              Alexandria, LA 25490 152200 0.09152   0.03790 0.01134
##   management
## 1         NA
## 2   0.054310
## 3         NA
## 4         NA
## 5   0.006509
## 6   0.015210
```

Let's add a new column, which records the *total* GMP, by multiplying the output per person by the number of people:

```
bea$gmp <- bea$pcgmp * bea$pop
```

And now let's look at this visually:

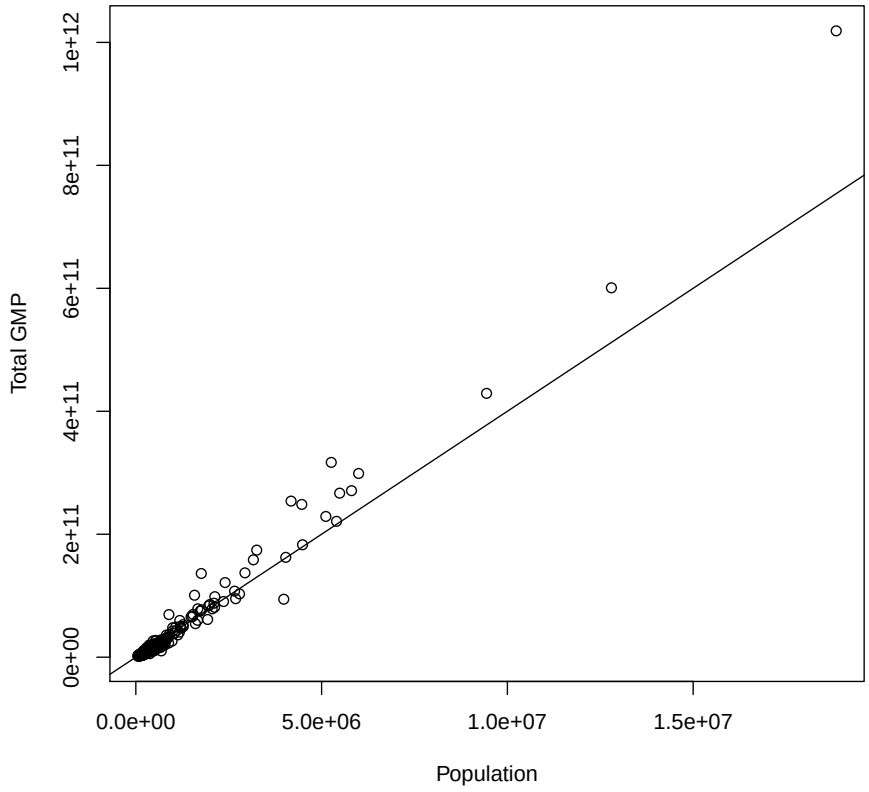
```
plot(gmp ~ pop, data = bea, xlab = "Population", ylab = "Total GMP")
```



The `plot` command, naturally enough, makes plots. The first argument to it tells it what to plot: here, we're telling it to plot `gmp` as a function of `pop`. (The tilde sign, is used in such "formulas" to indicate that what goes on the left is being treated as a function of what's on the right.) The next argument tells R where to look up the variables in the formula: in the data frame `bea` that we just loaded. The other two arguments give the axis some sensible labels.

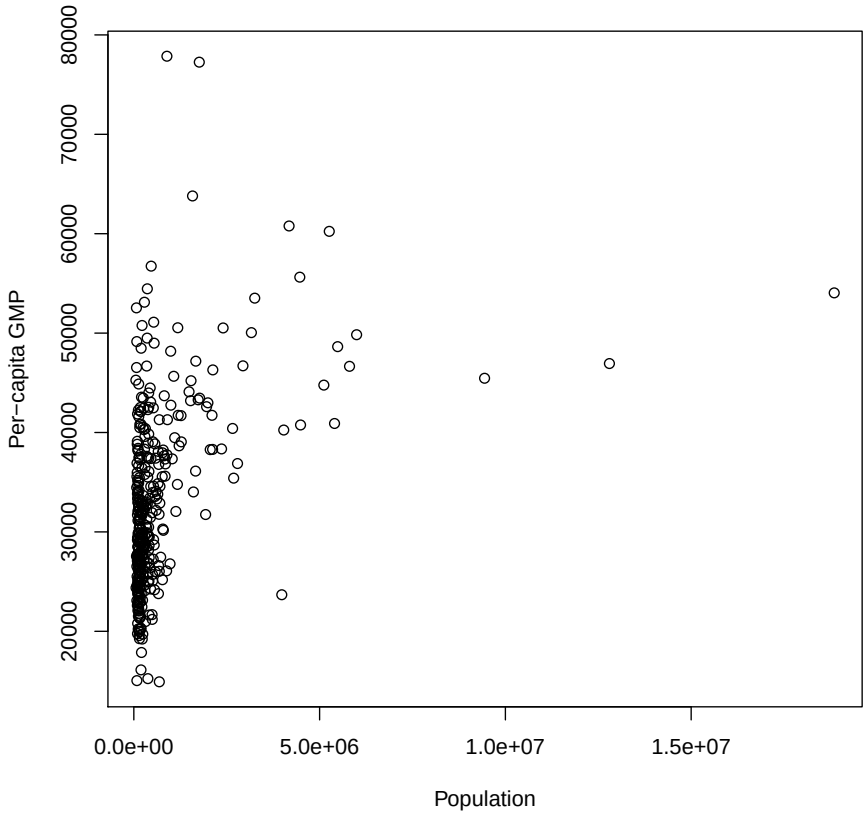
This plot shows, unsurprisingly, that larger cities have larger total economic outputs. What is more remarkable is how closely the points fall around a straight line. To see this, we'll re-plot the points, and use the function `abline` to add a straight line. To do this we need an intercept (the `a`) and a slope (the `b`). A reasonable guess for the intercept is 0 (since presumably a city with no inhabitants has no economy). One could reasonably guess the slope at 4×10^4 dollars/person, say by noticing the city with a population of about ten million (as it happens, Chicago) and seeing where it falls on the vertical axis.

```
plot(gmp ~ pop, data = bea, xlab = "Population", ylab = "Total GMP")
abline(a = 0, b = 40000)
```



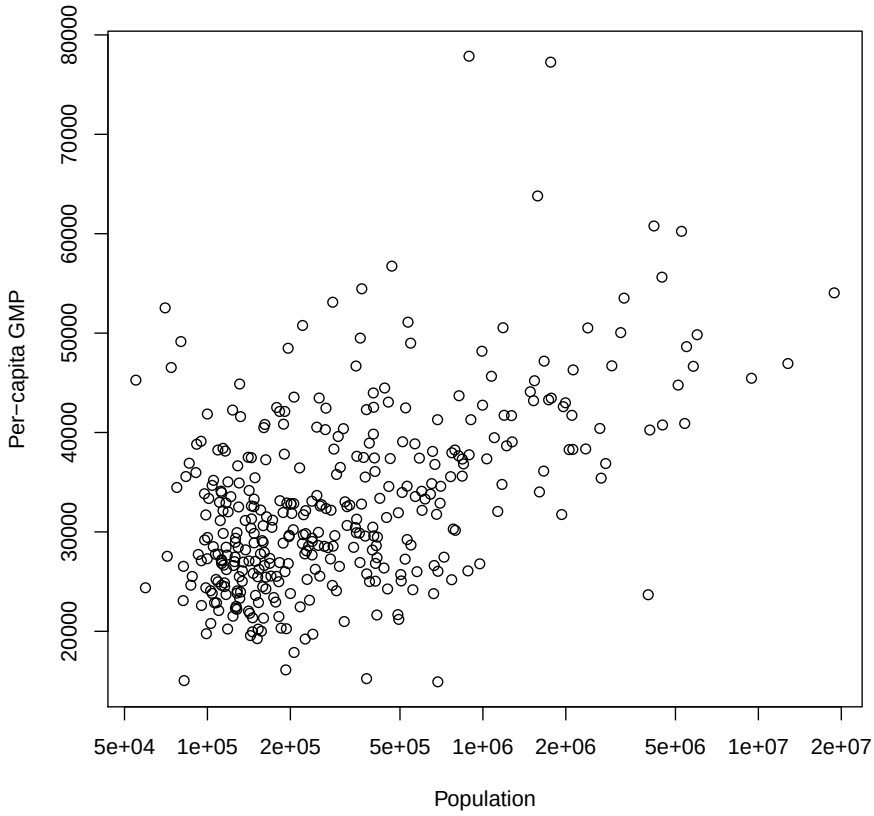
This isn't bad at all, but it looks like it's systematically too low for the larger cities. This is *suggestive* that there may be differences between the economies of large and small cities. Let's explore this by looking at the per-capita figures.

```
plot(pcgmp ~ pop, data = bea, xlab = "Population", ylab = "Per-capita GMP")
```



At this point, it becomes annoying that the larger cities in the US are so much larger than the small ones. By using a linear scale for the horizontal axis, we devote most of the plot to empty space around New York, Los Angeles and Chicago, which makes it harder to see if there is any trend. A useful trick is to switch to a logarithmic scale for that axis, where equal distances correspond to equal *multiples* of population.

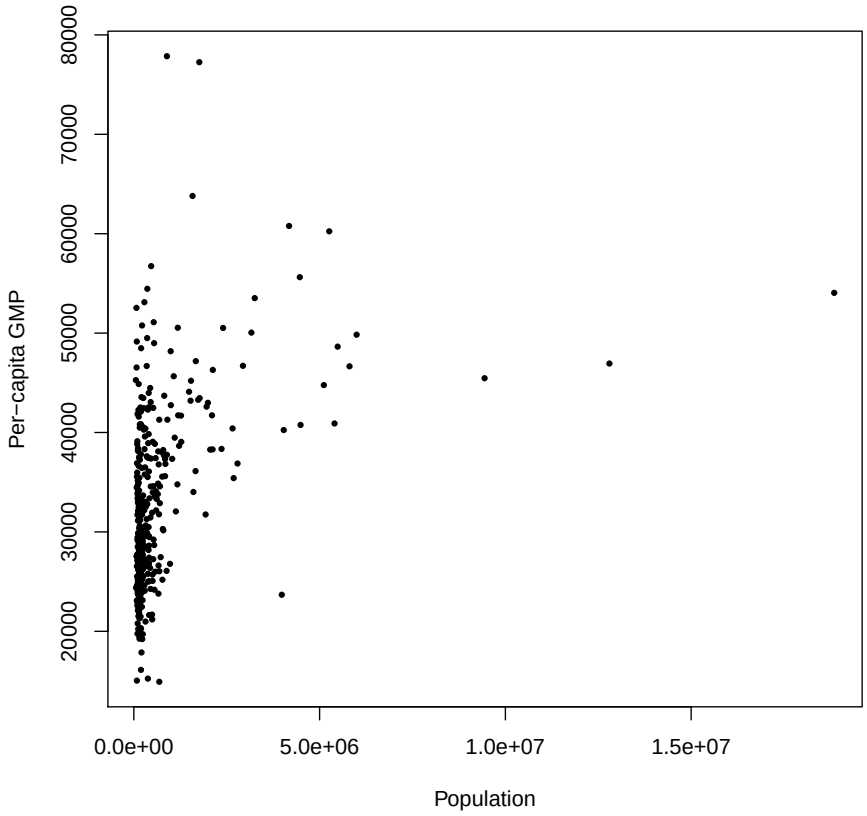
```
plot(pcgmp ~ pop, data = bea, xlab = "Population", ylab = "Per-capita GMP",
    log = "x")
```



Two things are noticeable from this plot. First, there is a wide range of per-capita GMP for smaller cities, which narrows as population grows. Second, there seems to be an increasing trend, or at least an increasing lower limit.

Let's restore the previous plot, but make it a bit less visually cluttered.

```
plot(pcgmp ~ pop, data = bea, xlab = "Population", ylab = "Per-capita GMP",
    pch = 19, cex = 0.5)
```



Let's now calculate our first regression line. R has a function for estimating linear models, with which we'll become very familiar:

```
lm(pcgmp ~ pop, data = bea)
##
## Call:
## lm(formula = pcgmp ~ pop, data = bea)
##
## Coefficients:
## (Intercept)          pop
##   3.128e+04    2.416e-03
```

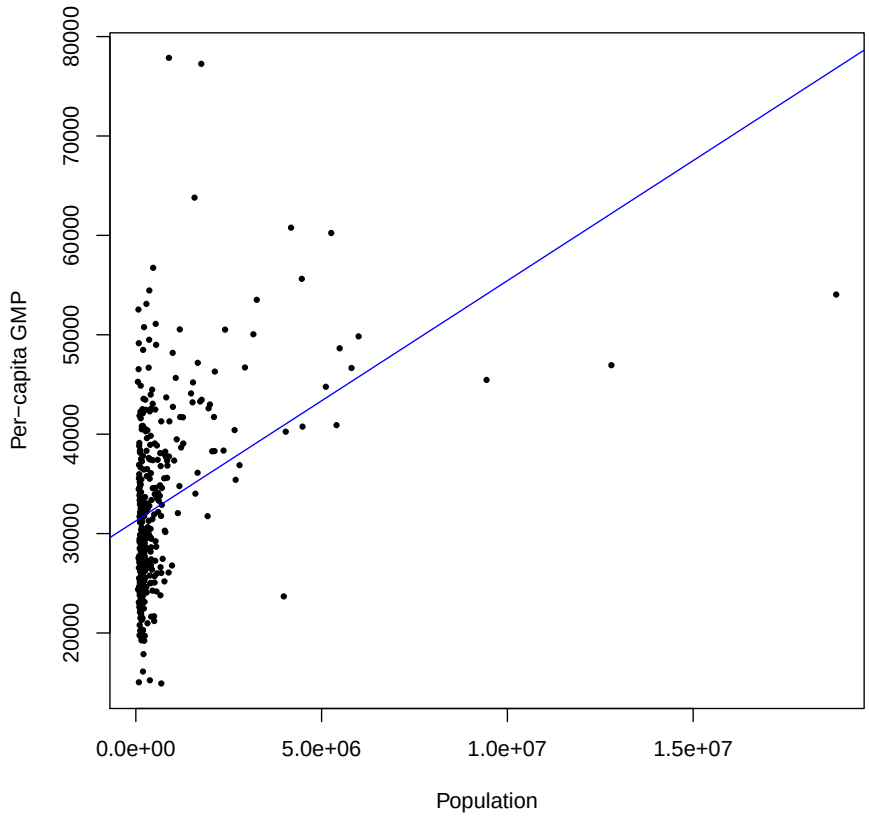
The first argument to `lm` is a formula, telling R which variable we're trying to predict (the one on the left, here `pcgmp`), and which variable we're trying to predict it from (the one on the right, here `pop`), and what data set to take those variables from (`bea` again)[1] R then *estimates* the coefficients of the best linear

[1]Much more complicated formulas are possible, but this will do for now.

predictor — we will see later how it does this — and returns those coefficients, along with a lot of other stuff which is invisible in this view.

The `abline` command is smart enough to get an intercept and a slope from the output of `lm`, so we can use it to decorate the plot:

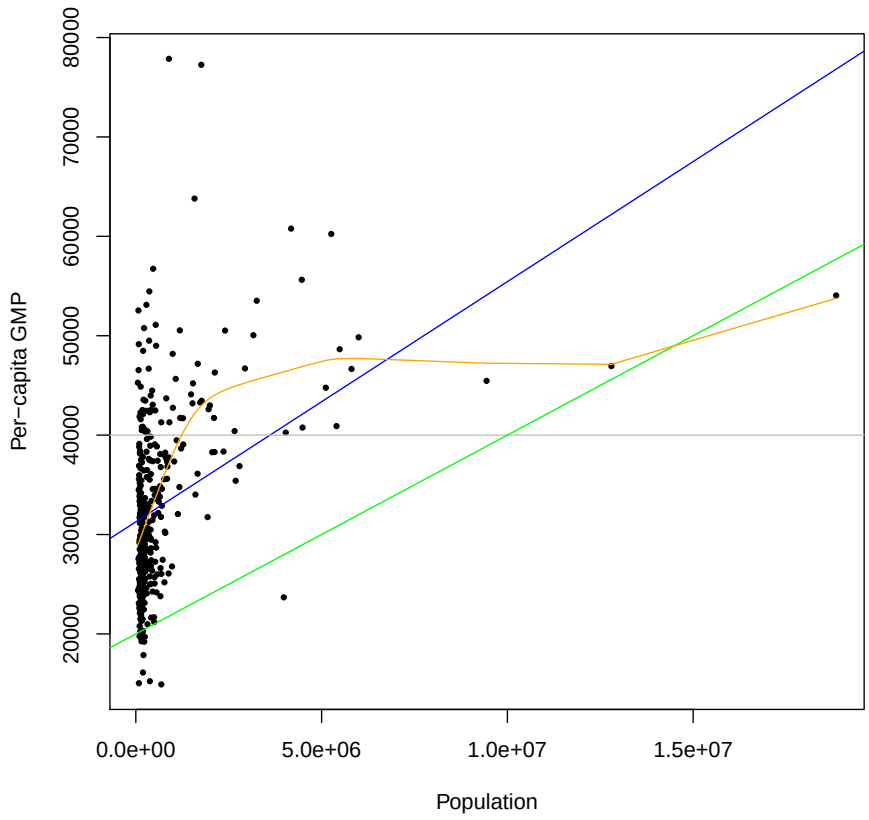
```
plot(pcgmp ~ pop, data = bea, xlab = "Population", ylab = "Per-capita GMP",
    pch = 19, cex = 0.5)
abline(lm(pcgmp ~ pop, data = bea), col = "blue")
```



But why should we believe that line? R does it, but I hope by this point in your life you don't think "the computer says it, so it must be right" is ever a good idea. Why prefer that blue line over this grey line, or the green one, or for that matter over the orange curve?

```
plot(pcgmp ~ pop, data = bea, xlab = "Population", ylab = "Per-capita GMP",
    pch = 19, cex = 0.5)
abline(lm(pcgmp ~ pop, data = bea), col = "blue")
abline(a = 40000, b = 0, col = "grey")
```

```
abline(a = 20000, b = 40000/2e+07, col = "green")
lines(smooth.spline(x = bea$pop, y = bea$pcgmp, cv = TRUE), col = "orange")
```



Why do we want to fit any lines here at all?

2 What Are Statistical Models For?

One basic use for these lines is as **summaries**. There is a lot of detail in those figures — each one requires $366 \times 2 = 732$ numbers. This is a lot of information to keep track of; it makes our heads hurt. In many situations, we'd rather ignore all of that precise detail and get away with a summary; we might want to *compress* the 732 numbers into just an intercept and a slope that describe the general trend or over-all shape of the data. There would be lots of ways to do that, and which ones would make sense would depend on how we want to use the compressed summary. We might, for instance, aim at being able to recover the original data with minimal error from the summary, so we'd want a

line which came close to the original points. But we might also decide that it was more important for the line to come closer to large cities than small ones (since they contain so many more people), etc., etc.

There is nothing wrong with data compression — it is, in fact, one of the fundamental technologies of modern life — but I have little more to say about it (here). The one thing I will say is that *anything* you can calculate from the data could, in principle, be used as a summary. Even if the calculation was originally inspired by doing some sort of statistical inference, every statistic can be a descriptive statistic.

If we want to go beyond describing, summarizing or compressing the data, we enter the realm of **inference** — we try to reach out and extend our knowledge from the data we have, to other variables we have not measured, or not measured so directly. This is inherently somewhat risky, imprecise, and uncertain. In statistics, we aim not only to draw such inferences, but to say something about the level of risk, imprecision, and uncertainty which accompanies them.

You have, by this point in your educations, been thoroughly schooled in one way in which inferences can be subject to uncertainty: when the data is just a sample of a larger population, and we want to extrapolate from the sample to the population. In fact, many people get taught that this is the only sort of uncertainty statistics can handle.

If that were true, there would be no role for statistics in dealing with this data set. It isn't any sort of sample at all — every city in the US in 2006 really is in there. So why, then, does it seem wrong to say that the slope of the optimal linear predictor is *exactly* 3.1277574×10^4, 0.0024162? There are at least two reasons.

One reason is that while we have measurements on the complete population, those measurements are themselves subject to error. In this case, while the people at the BEA try very hard to provide reliable numbers, their figures are the result of a complicated process, at which error can creep in at many points, from mistakes, accidents, deliberate lies, possibly-incorrect assumptions made at various stages, the use of random sampling in some steps, etc., etc.[2] Generally, just about every process of measurement is subject to some error: there are flaws in the measurement instrument or process, or we measure some imperfect proxy for what we're really interested in, or the measurement is perturbed by unrepeatable, irrelevant disturbances. In fact, much of the theory of mathematical statistics generally, and linear models specifically, was developed in the 19th century by astronomers and other physical scientists to quantify, and where possible reduce, the impacts of measurement error.

Some sources of measurement error are **systematic**: they tend to distort the measurement in predictable, repeatable ways. They may make the measured value larger than it should be, or smaller, or might shrink extreme values towards more mediocre ones, etc. Ideally, these systematic errors should be identified and modeled, so we can adjust for them. Other sources of error are (purely or merely)

[2]This is, among other things, a reason to want to compress the data, rather than just memorizing it: some part of the data is just wrong.

statistical: they are directionless and do not repeat, but the *distribution* of errors is predictable and stable; we can handle them with probability theory.

A second reason why it's reasonable to do statistical inference on a complete data set is that even the real values re somewhat accidental, and we'd like to know the general, underlying trends or relationships. Take the BEA data again: the *exact* values of the GMPs of all American cities in 2006 were to some extent the results of accidents, of chance, unrepeatable causes, and this would still be true even if we could measure those exact GMPs without error. To be really concrete for a moment, the city with the highest per-capita income in the data set is Bridgeport-Stamford-Norwalk, CT. This is a center for insurance companies, banks and hedge funds. Exactly how much money they made in 2006 depended on things like just how many hurricanes there were in Florida, wild fires in California, mortgages from Cleveland and Phoenix sold to small towns in Germany, etc., etc. Even if one thinks that there is, ultimately, some sort of deterministic explanation for all of these quantities, they're plainly very far removed from any general relationship between a city's population and its economic productivity. They really happen — they are not just measurement errors[3] — but they could easily have happened a bit differently. Long experience has shown that the *distribution* of these accidents is often stable, and can be modeled with probability theory[4]. When that happens, they are often called by the more respectable name of **fluctuations**.

To sum up: whether it's due to sampling, or measurement error, or fluctuations, we often have good reason to think that our data could have been more or less different. If we re-ran the experiment, or "re-wound the tape of history" (S. J. Gould), the results would not have been quite the same. This means that any statistic we calculate from the data would have been more or less different as well. When we try to quantify uncertainty, we want to know how different our calculations could have been.

To say anything useful here, we will need to make assumptions. Without *some* assumptions, we can't really say anything at all about how different the data could, plausibly, have been. In statistics, a lot of accumulated experience says that useful assumptions generally take the form of **statistical models** or **probability models**.

In a statistical model, we act as though the variables we measure, and possibly others we don't measure, are random variables. (We "model them as random variables.") The **specification** of a statistical model says what the random variables are, and lays down more or less narrow restrictions on their distributions and how they relate to each other. Here, for instance, are some *conceivable* statistical models for the BEA data. In all of them, X stands for a

[3] "As I regularly find myself having to remind cadet risk managers with newly-minted PhDs in financial econometrics, the Great Depression did actually happen; it wasn't just a particularly innaccurate observation of the underlying 4% rate of return on equities." (`http://d-squareddigest.blogspot.com/2006/09/tail-events-phrase-considered-harmful.html`)

[4] In fact, there are precise mathematical senses in which sufficiently complicated deterministic processes end up looking just like random ones. If this intrigues you, see **?** and **?**.

city's population and Y for its per-capita GMP.

1. $X \sim N(6.81 \times 10^5, 2.42 \times 10^{12})$; $Y|X \sim N(4.00 \times 10^4, 8.50 \times 10^7)$; X independent across cities; Y independent across cities given their X's.
2. $X \sim N(\mu_X, \sigma^2_X)$ for some mean μ_X and variance σ^2_X; $Y|X \sim N(4.00 \times 10^4, 8.50 \times 10^7)$; Y independent across cities given their X's.
3. $X \sim N(\mu_X, \sigma^2_X)$ for some mean μ_X and variance σ^2_X; $Y|X \sim N(4.00 \times 10^4, \sigma^2_Y)$ for some variance σ^2_Y; Y independent across cities given their X's.
4. distribution of X unspecified; $Y|X \sim N(4.00 \times 10^4, \sigma^2_Y)$ for some σ^2_Y; Y independent across cities given their X's.
5. distribution of X unspecified; $Y|X \sim N(\beta_0 + \beta_1 x, \sigma^2_Y)$ for some $\beta_0, \beta_1, \sigma^2_Y$; Y independent across cities given their X's.
6. distribution of X unspecified; $\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x$ for some β_0 and β_1; $\mathrm{Var}[Y|X = x] = \sigma^2_Y$ for some σ^2_Y; Y independent across cities given their X's.
7. distribution of X unspecified; $\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x$ for some β_0 and β_1; Y uncorrelated across cities given their X's.

As we go down this list of models, we make weaker and weaker assumptions about the process which generated the data. This means that, with the same data, we can infer less and less about that data-generating process. The very first model specifies a single, complete, unambiguous distribution for the data. If we assumed that model was true, we could then make all sorts of assertions about the distribution of city population and economic output, without even having to look at the data at all. Later models in the list leave more about the data-generating process undetermined, so we can use the data to estimate those parts of the model (e.g., the mean and variance of city populations in model 2, or the slope β_1 in models from 4 on), or test hypotheses about them, etc. Because a model specifies how the data variable X and Y are distributed, and any statistics we calculate are functions of X and Y, a model also, implicitly, tells us how those statistics are distributed[5]. These distributions of the statistics are what we'll use to quantify the uncertainty in our inferences.

The stronger the assumptions we make, the stronger the inferences we can draw, *if the assumptions are true*. There is no virtue to strong conclusions which rest on faulty premises. Therefore: we are going to go over how to draw these inferences, but we are also going to go over checking model assumptions.

When confronting a data analysis problem, you first need to **formulate** a statistical model. This is partly about absorbing what's already known about

[5]Whether we can work out that distribution in a nice closed form is another question, but mathematically exists, and there are ways to cope when there's no closed form, as we'll see later in this class, and in greater detail in 402.

the subject[6], partly about looking for similar problems others have dealt with and seeing what you can learn from them (i.e., analogy and tradition), and partly about being inspired by initial explorations of the data. Once you have a model, the two key tasks are **inference** within the model, and **checking** the model.

Inference within the model is about doing calculations which presume the model's assumptions are true: estimation, or prediction, or hypothesis testing, or confidence intervals, or what-have-you. This is usually what people actually want out of the data analysis. Unfortunately, these inferences are only as good as the modeling assumptions that they're based on.

Model checking, on the other hand, is about seeing whether the assumptions are really true. This includes formal goodness-of-fit testing, but also various "specification tests" (some of which we will cover), the less formal checks called "diagnostics", and sheer judgment, often aided by visualizations. If the assumptions of the model we started with turn out to be wrong, we need to go back and revise the model, either replacing the faulty assumptions with others, or weakening them.

People usually list put within-model inference before model checking — that's the order our textbook uses — but that's more because teachers, and students, are generally more comfortable with the more cut-and-dried topic of inference. That topic is extremely formalized and mathematical, with lots of theory to guide us. In fact, for lots of inference problems there is an unambiguous optimal procedure, which we should follow. Model checking, on the other hand, is much less formalized, mathematical and algorithmic than inference, and very little about it can be said to be definitely optimal or The Right Way To Do It. Nonetheless, *assumption checking is much more important*. Impressive-seeming inferences from strong-but-wrong assumptions don't actually tell us anything about the world, and are useless, no matter how much technical skill they might demonstrate. When reading other people's data analyses, you should get into the habit of paying very close attention to how they check their models, and you should apply that same habit to yourself.

3 The Simple Linear Regression Model

To make this philosophizing a bit more concrete, let's introduce the most basic of all statistical models that is actually useful for anything, the **simple linear regression model**. This is a model with two random variables, X and Y, where we are trying to predict Y from X. Here are the model's assumptions:

1. The distribution of X is unspecified, possibly even deterministic;
2. $Y|X = \beta_0 + \beta_1 x + \epsilon$, where ϵ is a noise variable;

[6] For instance, economists and geographers have long known about an "urban wage premium", where otherwise-similar workers in bigger cities get paid more (**?**).

3. ϵ has mean 0, a constant variance σ^2, and is uncorrelated with X and uncorrelated across observations.

The noise variable may represent measurement error, or fluctuations in Y, or some combination of both. The assumption of **additive** noise is non-trivial — it's not absurd to imagine that either measurement error or fluctuations might change Y multiplicatively (for instance). The assumption of a **linear functional form** for the relationship between Y and X is non-trivial; lots of non-linear relationships actually exist. The assumption of **constant variance**, or **homoskedasticity**, is non-trivial; the non-correlation assumptions are non-trivial. But the assumption that the noise has mean 0 *is* trivial. (Why?) Ideally, all of the non-trivial assumptions will be checked, and we will talk later in the course about ways to check them.

The assumptions I have just laid out, while they are non-trivial because they could be violated (and are, in many situations), are still strong enough to let us get a start on inference. While we will go into this in some detail next time, let's give this at least a start here.

Remember we saw last time that the optimal linear predictor of Y from X has slope $\beta_1 = \mathrm{Cov}[X,Y]/\mathrm{Var}[X]$. But both $\mathrm{Cov}[X,Y]$ and $\mathrm{Var}[X]$ are functions of the true distribution. Rather than having that full distribution, we merely have data points, say $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$. How might we *estimate* β_1 from this data?

An obvious approach would be to use the data to find the *sample* covariance and *sample* variance, and take their ratio. As a reminder, the sample variance of X is

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

while the sample covariance is

$$c_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

(Here I am writing $\overline{x}$ for the sample average of the x_i, and similarly for other variables.)[7] So we'd have a **sample** (or **empirical**) slope

$$\widehat{\beta_1} = \frac{c_{XY}}{s_X^2}$$

We can't hope that $\widehat{\beta_1} = \beta_1$, but we *can* hope that as $n \to \infty$, $\widehat{\beta_1} \to \beta_1$. When an estimator converges on the truth like that, the estimator is called **consistent**, and this is the most basic property a good estimator should have. What do we need to assume in order for $\widehat{\beta_1} \to \beta_1$?

[7] Some people prefer to define these with denominators $n-1$ rather than n, to get unbiased estimates of the population quantities. The way I am doing it will simplify some book-keeping presently.

Let's look at the sample covariance. A little algebra shows

$$c_{XY} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} \tag{1}$$

According to the model, $y_i = (\beta_0 + \beta_1 x_i + \epsilon_i)$. So (after a little more algebra)

$$\begin{aligned} c_{XY} &= \frac{1}{n}\sum_{i=1}^{n} x_i(\beta_0 + \beta_1 x_i + \epsilon_i) - \overline{x}\overline{\beta_0 + \beta_1 x + \epsilon} & (2)\\ &= \beta_0\overline{x} + \beta_1\overline{x^2} + \overline{x\epsilon} - \overline{x}\beta_0 - \beta_1\overline{x}^2 - \bar{x}\bar{\epsilon} & (3)\\ &= \beta_1 s_X^2 + \overline{x\epsilon} - \bar{x}\bar{\epsilon} & (4) \end{aligned}$$

Because ϵ has mean 0, as $n \to \infty$, the law of large numbers says $\bar{\epsilon} \to 0$. Because ϵ is uncorrelated with x, using the law of large numbers again says that $\overline{x\epsilon} \to 0$ as well. So

$$c_{XY} \to \beta_1 s_X^2$$

and therefore

$$\widehat{\beta_1} = \frac{c_{XY}}{s_X^2} \to \frac{\beta_1 s_X^2}{s_X^2} = \beta_1$$

as desired.

As I said, this argument rests on all the model assumptions. Strictly speaking, the estimator $\widehat{\beta}$ is consistent under even weaker assumptions — it's enough that $c_{XY} \to \mathrm{Cov}[X,Y]$, and $s_X^2 \to \mathrm{Var}[X]$. On the other hand, it would be nice to say more: we want to know *how far* from the truth our estimate is likely to be, whether it tends to over- or under- estimate the slope, etc. we will see in later lectures how the assumptions of the simple linear regression model will let us say something about all of these matters, and how the even stronger assumption that the noise is Gaussian will let us be even more precise.

(We will, I promise, come back to this data set, and the question of which regression line, if any, best describes the relationship between a city's size and its economic output, but that, too, will have to wait for later.)

Exercises

To think through, not to turn in.

1. What, if anything, makes `plot(pcgmp ~ log(pop), data=bea)` a worse plot than `plot(pcgmp ~ pop, data=bea, log="x")`?
2. Fill in the algebra for (1).
3. Fill in the algebra for (2).